Revista Electrónica de Metodología Aplicada
1999, Vol. 4 nº 1, pp. 1-8

LA PRUEBA CHI-CUADRADO EN TABLAS DE CONTINGENCIA CON CELDAS VACIAS: UN PROCEDIMIENTO EN SPSS.

Fco. Javier Herrero
Marcelino Cuesta
Paula Fernández
Dpto. de Psicología
Universidad de Oviedo
e-mail:herrero@correo.uniovi.es

ABSTRACT.

In this paper a procedure is presented, created in the statistical program SPSS, that allows the analysis of bidimensional contingency tables when the relatively frequent situation of the presence of empty cells takes place.The program follows the algorithm described by Graf, Alf and Williams (1997). It is also presented an example of use of the proposed tool.

Key words: Data Analysis, contingency tables, missing values, chi square, SPSS.

1. Introducción.

Frecuentemente en el campo de las Ciencias del Comportamiento, a la hora de analizar la asociación en una tabla de contingencia (m filas y n columnas), nos encontramos con celdas vacías. Esta falta de información produce el resultado desagradable de imposibilitar la obtención del estadístico de relación chi-cuadrado.

Una de las posibilidades de solventar el problema anterior, consistiría en reagrupar las categorías de las variables estudiadas, con objeto de eliminar las celdas vacías en la tabla de frecuencias. El problema que se suele plantear con esta alternativa, es que hacemos una reagrupación de la información que inicialmente no se observó de este modo, es decir, estamos provocando un artefacto en el análisis de los resultados.

Si tenemos muy en cuenta esta objeción, el procedimiento correcto sería utilizar un método que manteniendo la dimensionalidad inicial, nos permita también obtener la información sobre la asociación de las variables estudiadas.

En la literatura podemos encontrar distintas soluciones, como son los algoritmos propuestos por Savage y Deutsch (1960), Godman (1968) Mantel (1970) y Wagner (1970). De todas ellas, en este artículo, hemos seguido la solución de remplazar las celdas vacías por

los valores esperados (Graf, Alf y Willians, 1997) que corresponderían en el caso de cumplirse la hipótesis nula, es decir la ausencia de relación entre las dos variables. La ventaja de este algoritmo, es que la información contenida en las celdas vacías no implican ninguna contribución en el resultado final del estadístico de significación.

2.- Algoritmo general de resolución.

Consideremos la formulación clásica de relación entre dos variables, cuya información se ha recogido en una tabla de contingencias(mxn):

$$\chi^2=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(fo_{ij}-fe_{ij})^2}{fe_{ij}} \qquad \textbf{(1)}$$

donde…
m = número de filas de la tabla de frecuencias
n = número de columnas de la tabla
fo_{ij} = frecuencia observada para la fila i columna j de la tabla de frecuencias.
fe_{ij} = frecuencia teórica(esperada) para la fila i columna j de la tabla de frecuencias.

y siendo…

$$fe_{ij}=\frac{f_{i.}\cdot f_{.j}}{N} \qquad \textbf{(2)}$$

donde …
N = número total de observaciones en la tabla
$f_{i.}$ = número de observaciones en la fila i
$f_{.j}$ = número de observaciones en la columna j

Si tenemos alguna celda vacía, al estar la información incompleta no es posible obtener el estadístico de significación (fórmula 1), ni tampoco algunos de los términos de las frecuencias esperadas (fórmula 2).

Supongamos, entonces que nos encontramos en esta situación. Tenemos una celda vacía ($f_{ij} = 0$), y que a la frecuencia de esa celda vacía la denotamos como fv_{ij}. Entonces, deberemos sustituir la celda vacía (fv_{ij}) por un valor, que en este algoritmo consiste en una puntuación esperada (fe_{ij}) que no va a tener contribución significativa en el estadístico final.

De este modo la tabla de contingencias sufre la siguiente transformación:
$N^* = N + fv_{ij}$ **(3)**
$f^*_{i.} = f_{i.} + fv_{ij}$ **(4)**
$f^*_{.j} = f_{.j} + fv_{ij}$ **(5)**

donde …
N* = número total de observaciones en la tabla incluido las celdas vacías transformadas.
$f^*_{i.}$ = número de observaciones en la fila i con celdas vacías transformadas
$f^*_{.j}$ = número de observaciones en la columna j con celdas vacías transformadas

El siguiente paso para estimar celda vacía (fv_{ij}), suponiendo la independencia de filas y columnas, consistirá en sustituir las ecuaciones (3, 4, 5), resultando:

$$fv_{ij} = \frac{(f_{i.} + fv_{ij})(f_{.j} + fv_{ij})}{N + fv_{ij}} \qquad \textbf{(6)}$$

resolviendo (fv_{ij}) en (6) obtendremos:

$$fv_{ij} = \frac{f_{i.} \cdot f_{.j}}{N - f_{i.} - f_{.j}} \qquad \textbf{(7)}$$

Vamos a ver un ejemplo numérico de lo dicho hasta ahora, supongamos que la tabla de frecuencias es la siguiente (una sola celda vacía):

Tabla de contingencia FILA * COLUMNA

Recuento

		COLUMNA			
		1.00	2.00	3.00	Total
FILA	1.00		2	3	5
	2.00	4	5	6	15
Total		4	7	9	20

Tabla I: Frecuencias observadas originales.

Operando la fórmula anterior (7), el valor de celda 1,1 sería:

$$fv_{11} = \frac{5 \cdot 4}{20 - 5 - 4} = 1{,}82$$

que será asignado entonces a la celda vacía anterior. Este valor sería utilizado en las expresiones (1) y (2), con lo cual podríamos obtener el estadístico χ^2. Teniendo en cuenta que los grados de libertad deberán ser corregidos de acuerdo a la siguiente expresión:

$gl = (m-1)(n-1)-1$
debido a que hemos perdido un grado de libertad a la hora de estimar la celda (fv_{11}).

Por desgracia, la ecuación (7), no se puede aplicar de forma directa cuando el número de celdas vacías son más de una. Lo más frecuente en estas situaciones es que la solución sea desconocida. La forma de resolverlo es utilizar un método iterativo, como el propuesto por Scarborough (1966).

3.- Procedimiento en SPSS: un ejemplo.

La mejor forma de solucionar los problemas de cálculo, sobre todo cuando tenemos más de una celda vacía es desarrollar un procedimiento mecanizado. En este artículo se describe la implementación en SPSS. Para ello hemos utilizado el procedimiento general MATRIX, en el cual hemos implementado el algoritmo general descrito por Scarborough (1966). Se ha procurado facilitar su manejo en lo posible, de tal forma que solo el usuario debe introducir la tabla de contingencias, y después ejecutar el procedimiento para llegar a la solución deseada.

Supongamos que deseamos analizar la siguiente tabla de contingencias, donde se intenta establecer la asociación entre la clase social del individuo y su rendimiento académico:

Tabla de contingencia Clase Social * Rendimiento académico

Recuento

		Rendimiento académico			
		Bajo	Medio	Alto	Total
Clase Social	Clase Baja	14	6		20
	Clase Media	9	15	6	30
	Clase Alta		4	6	10
Total		23	25	12	60

Tabla II: Frecuencias observadas originales.

Como se puede comprobar en la tabla anterior existen dos celdas vacías, lo cual imposibilita la obtención del estadístico clásico de χ^2.

El procedimiento implementado por nosotros en SPSS (ver apéndice), es muy fácil de utilizar. Es suficiente sustituir la información recogida en la tabla II en el programa de la siguiente forma:

```
compute X={14, 6, 0;
           9,15, 6;
           0, 4, 6}.
```

Tabla III: Implementacion de la tabla de contingencias.

Y a continuación ejecutar el procedimiento, que nos dará como resultado:

```
Test de Chi-Cuadrado:
  Chi-Cuad        GL       Sig.
    6.6051    2.0000     .0368
```

Tabla IV: Estadísticos de significación.

Teniendo en cuenta que los grados de libertad para este caso se obtienen a partir de la expresión:

$$gl = (m-1)(n-1)-k$$

donde k es el número de celdas vacías.

4.- Conclusiones.

El programa descrito, que permite estimar las celdas vacías en una tabla de contingencias, puede ser como hemos podido ver anteriormente un procedimiento muy útil en las Ciencias del Comportamiento, donde con frecuencia nos encontramos con falta la información en alguna de las celdas de las tablas de contingencia, sobre todo si la dimensionalidad es alta y trabajamos con muestras reducidas.

El método es útil siempre que se aplique a tablas con dimensiones superior a 2x2, y además donde el número de celdas vacías no sea muy elevado, ya que en otro caso los grados de libertad imposibilitarían una solución correcta. Nótese que los grados de libertad en el procedimiento descrito varían en función de las dimensiones de la tabla y del número de celdas vacías, lo cual produce una disminución gradual de los grados de libertad en la medida que se incrementan los valores perdidos.

5.- Referencias.

Godman, L.A. (1968). The analysis of cross-classified data: Independence, quasi-independence, and interactions in contingency tables with o without missing values. *Journal of the American Statistical Association, 63*, 1091-1131.

Graf, R.G.; Alf, E.F.Jr. y William s, S. (1997). A computer program for estimating missing cell frequencies in chi square tests for association. *Interstat, Agosto*.

Mantel, N. (1970). Incomplete contingency tables. *Biometrics, 26*, 291-304.

Savage, I.R. y Deutsch, K.W. (1960). A statistical model of the gross analysis of transation flows. *Econometrica, 28*, 551-572.

Scarborough, J.B. (1996). *Numerical Mathematical Analysis*. Baltimore: The Johns Hopkins Press.

Wagner, S.S. (1970). The Maximum-Likelihood estimate for contingency tables with zero diagonal. *Journal of the American Statistical Association, 65,* 1362-1383.

6.- Apéndice.

El siguiente procedimiento está escrito para SPSS para Windows. Y ha sido comprobado su funcionamiento en la versión SPSS(8.0) para Windows.

```
*Determinación de la asociación en tablas de contingencia
*con celdas nulas o missing.

SET MXLOOPS=199 MITERATE 199.
Matrix.


*Parámetros a modificar por el usuario.
*=====================================.
*Definición de la tabla de contingencias que se
* desea analizar.
* Se deberá sustituir los codigos de las celdas
* por los valores empíricos.
compute X={c11, c12, c13;
           c21, c22, c23;
           c31, c32, c33}.
*Identificación del valor nulo o missing de la tabla.
compute nulo=0.
*-----------------------------------------.


*********************************.
*Algoritmo general de resolución.
*_______________________________.
* Definición de los escalares.
compute m=nrow(X).
compute n=ncol(X).
compute k=0.
compute xnulo=0.
compute onulo=0.
compute chi2=0.

*Definición de Matrices y vectores.
compute O=make(m,n,0).
compute E=make(m,n,0).
compute vf=make(m,1,0).
compute vc=make(1,n,0).
compute vf2=make(m,1,0).
compute vc2=make(1,n,0).

* Determina el número de celdas nulas o missing.
+loop i=1 to m.
+ loop j=1 to n.
+   do if X(i,j)=nulo.
+   compute k=k+1.
+  end if.
+ end loop.
+end loop.
```

```
+loop i=1 to m.
+ loop j=1 to n.
+   do if X(i,j)<>nulo.
+   compute xnulo=xnulo+X(i,j).
+   compute vf(i)=vf(i)+X(i,j).
+   compute vc(j)=vc(j)+X(i,j).
+  end if.
+ end loop.
+end loop.

+loop i=1 to m.
+ loop j=1 to n.
+  do if X(i,j)<>nulo.
+   compute O(i,j)=X(i,j).
+  end if.
+ end loop.
+end loop.

compute itera=0.
+loop.
+ compute itera=itera+1.
+ compute difer=0.
+ loop i=1 to m.
+  loop j=1 to n.
+   do if X(i,j)=nulo.
+    compute f=(vf(i)-O(i,j))*(vc(j)-O(i,j)).
+    compute f=f/(xnulo-vf(i)-vc(j)+O(i,j)).
+    compute xnulo=xnulo-O(i,j)+f.
+    compute vf(i)=vf(i)-O(i,j)+f.
+    compute vc(j)=vc(j)-O(i,j)+f.
+    compute difer=difer+abs(O(i,j)-f).
+    compute O(i,j)=f.
+   end if.
+  end loop.
+ end loop.
+end loop if (difer<.0001).

print X/title="Matriz inicial".
print nulo/title="El valor siguiente expresa una celda nula o
missing".
print /title "----------------------------".
print O/title="Matriz trasformada"/format=F8.2.
print itera/title="Numero de iteraciones realizadas:".

+loop i=1 to m.
+ loop j=1 to n.
+  compute onulo=onulo+O(i,j).
+  compute vf2(i)=vf2(i)+O(i,j).
+  compute vc2(j)=vc2(j)+O(i,j).
+ end loop.
+end loop.

+loop i=1 to m.
+ loop j=1 to n.
+  compute E(i,j)=vf2(i)*vc2(j)/onulo.
```

```
+ end loop.
+end loop.

print E/title="Matriz de valores esperados"/format=f8.3.

+loop i=1 to m.
+ loop j=1 to n.
+  compute chi2=chi2+(O(i,j)-E(i,j))**2/E(i,j).
+ end loop.
+end loop.

compute gl=(m-1)*(n-1)-k.
compute p=1-CHICDF(chi2,gl).
compute test={chi2,gl,p}.
print /title="-----------------------------------------".
print test /format "f9.4"/title 'Test de Chi-Cuadrado:'
  /space 1/cnames={"Chi-Cuadrado","  GL  ","  Sig."}.

end matrix.
```